LA-UR-98-2776

Title: **Visualization of
High-Dimensional Clusters
Using Nonlinear Magnification**

Author(s): T. Alan Keahey

Submitted to: Conference on

Visual Data Exploration and Analysis

June 16, 1998

http://lib-www.lanl.gov/la-pubs/00412896.pdf

Los Alamos
N A T I O N A L   L A B O R A T O R Y

Visualization of High-Dimensional Clusters Using Nonlinear Magnification

T. Alan Keahey
Graphics and Visualization Team
CIC-8, MS B287
Los Alamos National Laboratory
Los Alamos, NM 87545 USA
(505) 667-3233
(505) 665-4939 (fax)
keahey@lanl.gov

June 26, 1998

1   Extended Abstract

This paper describes a cluster visualization system used for data-mining fraud detection. The system can simultaneously show 6 dimensions of data, and a unique technique of 3D nonlinear magnification allows individual clusters of data points to be magnified while still maintaining a view of the global context. We will first describe the fraud detection problem, along with the data which is to be visualized. Then we will describe general charateristics of the visualization system, and show how nonlinear magnification can be used in this system. Finally we conclude and describe options for further work.

1.1   Application: Medicare Fraud Detection

The cluster visualization system described in this paper was developed in 1996 as part of a data-mining project at Los Alamos National Laboratory, sponsored by the US Federal Health Care Finance Administration, for detecting fraud among Medicare providers and beneficiaries. The original data for examination is composed of N attributes (both categoric and numeric) for each data point (provider record). K-means analysis is then used to partition the N-dimensional space into 100 clusters, each represented by an N-dimensional cluster centroid. Associated with each of these centroids and data points is a *probability density function* (PDF) estimate which (informally stated) reflects the probability with which we would expect to find "normal" items within a given region of the N-dimensional space. For the specific examples in this

paper, the dataset is composed of $35,000$ 11-dimensional records for medicare providers from a single state in the southeastern United States.

1.2 Cluster Visualization

Because of the relative sparsity of the cluster data, it is possible to effectively lay out more than three dimensions of the data in a single 3D coordinate system. For a data set composed of N-dimensional points (each represented by an N-tuple $\{x_1, x_2, \ldots, x_n\}$) we can select *frames* of the data with each frame representing 3 of the N possible dimensions. Multiple frames can then be laid out within the 3D coordinate space of the visualization, using colour cues to visually separate the dimensions. As an example, one frame could be composed of the dimensions $\{x_1, x_2, x_3\}$ and be rendered in green, while another frame could have the dimensions $\{x_5, x_4, x_8\}$ and be rendered in blue. Figure 1 shows a screen snapshot of the program, with 6 dimensions of the data being rendered.

Individual records are rendered as single points, with transparency proportional to the PDF estimate so that unusual records are more clearly visible. Cluster centroids are rendered as wireframe boxes, with box size being inversely proportional to the PDF estimate. Nonlinear scaling of the PDF values (similar to gamma-correction methods) can be used in conjunction with explicit clipping of items based on PDF values to selectively render only those items having a given PDF value or lower, thus allowing the user to better focus on items of interest.

Selection of data records can either be done by entering the record ID number, or by selecting one or more data points with the mouse. When a record is selected, it is highlighted in all of the frames on-screen to show it's position relative to the other records in the visible dimensions. Once a record is selected, the user can highlight the cluster centroid associated with the record, and once a cluster centroid is selected, all of the data records associated with that cluster can be highlighted.
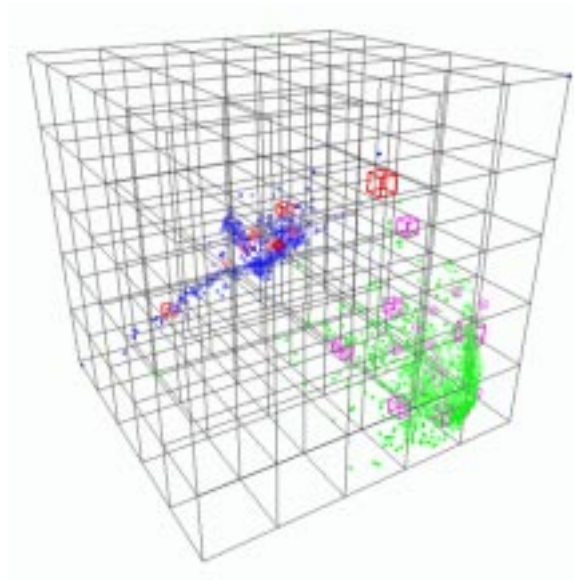

Figure 1: Cluster Visualization

1.3   Nonlinear Magnification

Many approaches have been described in the literature for stretching and distorting spaces to produce effective visualizations. The term *nonlinear magnification* was introduced in [2] to describe the effects common to all of these approaches.The basic properties of nonlinear magnification are non-occluding in-place magnification which preserves a view of the global context. Most of the existing nonlinear magnification systems to-date have involved the magnification of 2D information spaces. Many of these systems also rely on perspective projections of a mapping of the 2D information space onto a 3D manifold in order to create the magnification effect [4, 6], and are thus constrained to the magnification of 2D spaces only. In contrast to these perspective-based systems, transformation-based techniques such as the nonlinear transformations of [2] and the hyperbolic spaces of [5] allow for simple and direct extension to 3 or more dimensions. This visualization system uses 3D versions of the transformations described in [2] simply by the addition of an additional z coordinate which is treated similarly to the x and yy coordinates.

Visualization in three dimensions inherently involves occlusion of some portions of the data, however many methods for dealing with occlusion are available, such as clipping, transparency-based rendering, and creating "tunnels" through the data in order to see an occluded point of interest [1]. Occlusion does not present a great problem for this cluster visualization application however, as the data is composed of dense clusters in a relatively sparse space, and there is a great deal of empty space available within the information space in which to perform the magnification. Other likely candidates for 3D nonlinear magnification with similar sparsity properties might involve graph visualization as in [5].

The techniques presented in [2] describe many different magnification effects that can be acheived using combinations of simple and computationally efficient 2D transformation functions. All of these effects have a straightforward extension to 3D viewing, although we illustrate only a few of them here in Figure 2. The leftmost figure illustrates the use of an unbound orthogonal transformation to expand a single cluster of data. The figure on the right shows the combination with two independent centers of magnification; each magnification center uses a bounded radial transformation to expand an individual cluster of data within the overall space.

1.4   Conclusions and Further Work

This cluster-visualization application combines several different techniques to enhance visualization for data mining. The use of frames allows high-dimensional visualization, while the transparency based rendering helps to reduce visual clutter to focus on the more important items of interest. Nonlinear magnification is also employed to enhance the view of one or more clusters while simultaneously allowing a view of the global context.

There are several potential areas for further research and improvements to the prototype visualizer. Implicit magnification fields were introduced in [3] as a method for determining the amount of magnification that is implicit in complex transformations of the type described here. By synchronizing rendering functions to this implicit magnification field, significant efficiency gains may become possible if data points below a certain magnification level are aggregated or eliminated from the rendering. A method is also described in [3] that allows properties of the data to directly define spatial transformations as a field of scalar magnification values, these data-driven magnifications offer the potential to automatically provide visual enhancement of the regions which are of greatest interest to the user.
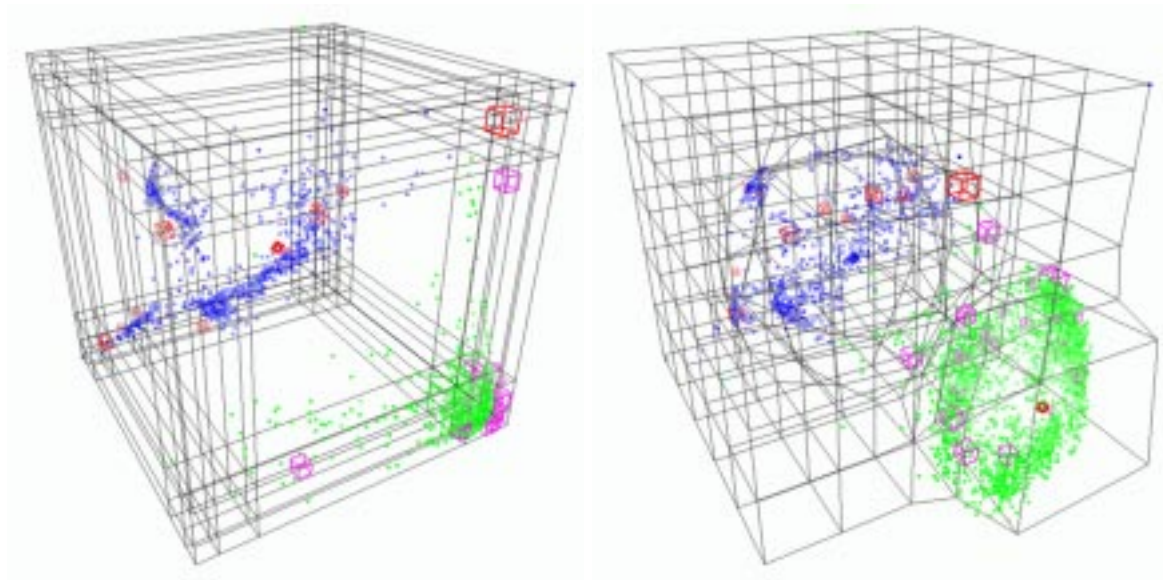
Figure 2: Using Unconstrained and Constrained Magnification

References

[1] M.S.T. Carpendale, D.J. Cowperthwaite, and F.D. Fracchia. Distortion viewing techniques for 3D data. In *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, pages 46–53, 1996.

[2] T. Alan Keahey and Edward L. Robertson. Techniques for non-linear magnification transformations. In *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, pages 38–45, October 1996.

[3] T. Alan Keahey and Edward L. Robertson. Nonlinear magnification fields. In *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, October 1997.

[4] J.D. Mackinlay, G.G. Robertson, and S.K. Card. The perspective wall: Detail and context smoothly integrated. In *Proceedings of the ACM Conference on Computer Human Interaction*, pages 173–179, 1991.

[5] Tamara Munzner. H3: Laying out large directed graphs in 3D hyperbolic space. In *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, October 1997.

[6] G. Robertson and J. D. Mackinlay. The document lens. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 101–108, 1993.

2 Keywords

visualization, nonlinear magnification, fisheye views, data mining

3 Biography

Alan Keahey received his B.C.Sc. (Honours) from the University of Manitoba in 1992, and his M.S. (1994) and Ph.D. (1998) degrees in Computer Science from Indiana University, Bloomington. He now works as a Visualization Research Scientist at Los Alamos National Laboratory. His research interests include scientific and information visualization, computer graphics, and human-computer interaction.